(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 September 2002 (19.09.2002)

PCT

(10) International Publication Number
WO 02/073463 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/SE02/00462

(22) International Filing Date: 13 March 2002 (13.03.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0100856-4 13 March 2001 (13.03.2001) SE

(71) Applicant *(for all designated States except US)*: **PICSEARCH AB** [SE/SE]; Hammarby Fabriksväg 23, S-120 33 Stockholm (SE).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RISBERG, Robert** [SE/SE]; Omgången 436-11, S-412 80 Göteborg (SE). **ANDERSSON, Nils** [SE/SE]; Torphagsvägen 14, SE-104 05 Stockholm (SE).

(74) Agents: **BERGLUND, Stefan** et al.; Bjerkéns Patentbyrå KB, Östermalmsgatan 58, S-114 50 Stockholm (SE).

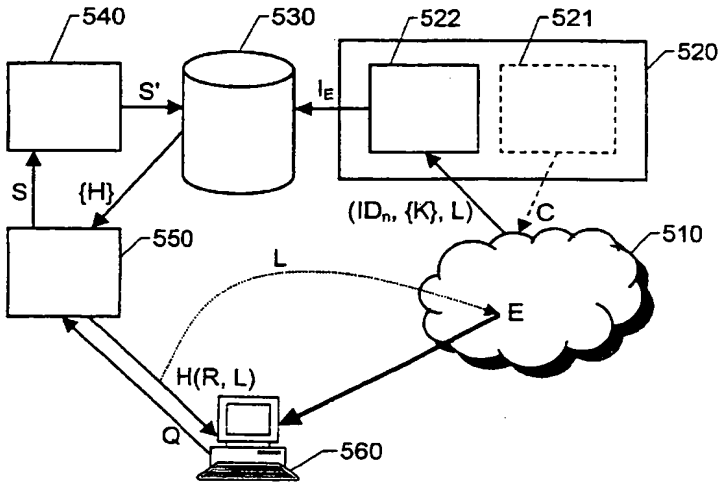
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: INDEXING OF DIGITISED ENTITIES



(57) Abstract: The invention relates to indexing of digitised entities (E) in a large and comparatively unstructure data collection (510), for insstance the Internet, such that text-based searches (S; S') wht respect to the data collection (510) can be ordered (Q) via a user client terminal (560). Index information (I_E) is generated (522) for each digitised entity (E), which contains distinctive features ({K}) being ranked according to a rank parameter. The rank parameter indicates a degree of relevanceof particular distinctive freature (K) wiht respect to a give digitised entity (E) and is derived from fields or tags associated wiht one or more copies of the digitised entity (E) in the data collection (510). The index information (I_E) is stored in a searchable database (530), which is accessible via a user client interface (550) and a serch engine (540). The derived distinctive features (K) and the rank parameter thus provides a possibility to carry out text-based searches (Q) in respect of non-text digitised entities (E), such as images, audio files and video sequences and obtain a highly relevant search result ({H}).

WO 02/073463 A1

Indexing of Digitised Entities

THE BACKGROUND OF THE INVENTION AND PRIOR ART

The present invention relates generally to indexing of digitised entities in a large and comparatively unstructured data collection such that a relevant search result can be obtained. More particularly the invention relates to a method of indexing digitised entities, such as images, video or audio files, according to the preamble of claim 1. The invention also relates to a computer program according to claim 13, a computer readable medium according to claim 14, a database according to claim 15 and a server/client system according to the preamble of claim 16.

Search engines and index databases for automatically finding information in digitised text banks have been known for decades. In recent years the rapid growth of the Internet has intensified the development in this area. Consequently, there are today many examples of very competent tools for finding text information in large and comparatively unstructured data collections or networks, such as the Internet.

As the use of the Internet has spread to a widened group of users, the content of web pages and other resources has diversified to include not only text, but also other types of digitised entities, like graphs, images, video sequences, audio sequences and various other types of graphical or acoustic files. An exceptionally wide range of data formats may represent these files. However, they all have one feature in common,

namely that they per se lack text information. Naturally, this fact renders a text search for the information difficult. Various attempts to solve this problem have nevertheless already been made.

For instance, the US patent 6,084,595 describes an indexing method for generating a searchable database from images, such that an image search engine can find content based information in images, which match a user's search query. Feature vectors are extracted from visual data in the images. Primitives, such as colour, texture and shape constitute parameters that can be distilled from the images. A feature vector is based on at least one such primitive. The feature vectors associated with the images are then stored in a feature database. When a query is submitted to the search engine, a query feature vector will be specified, as well as a distance threshold indicating the maximum distance that is of interest for the query. All images having feature vectors within that distance will be identified by the query. Additional information is computed from the feature vector being associated with each image, which can be used as a search index.

An alternative image and search retrieval system is disclosed in the international patent application WO99/22318. The system includes a search engine, which is coupled to an image analyser that in turn has access to a storage device. Feature modules define particular regions of an image and measurements to make on pixels within the defined region as well as any neighbouring regions. The feature modules thus specify parameters and characteristics which are important in a particular image match / search routine. As a result, a relatively rapid comparison of images is made possible.

The international patent application WO00/33575 describes a search engine for video and graphics. The document proposes the creation and storage of identifiers by searching an area within a web page near a graphic file or a video file for

searchable identification terms. Areas on web pages near links to graphic or video files are also searched for such identification terms. The identification terms found are then stored in a database with references to the corresponding graphic and video files. A user can find graphic or video files by performing a search in the database.

However, the search result will, in general, still not be of sufficiently high quality, because the identification terms are not accurate enough. Hence, relevant files may either end up comparatively far down in the hit list or be missed completely in the search.

SUMMARY OF THE INVENTION

It is therefore an object of the present invention to alleviate the problem above and thus provide an improved solution for finding relevant digitised entities, such as images, video files or audio files, by means of an automatic search being performed with respect to a large and relatively unstructured data collection, such as the Internet.

According to one aspect of the invention the object is achieved by a method of indexing digitised entities as initially described, which is characterised by generating index information for a particular digitised entity on basis of at least one rank parameter. The rank parameter is derived from basic information, which in turn pertains to at least one distinctive feature and at least one locator for each of the digitised entities. The rank parameter indicates a degree of relevance for at least one distinctive feature with respect to each digitised entity.

According to another aspect of the invention these objects are achieved by a computer program directly loadable into the internal memory of a digital computer, comprising software for controlling the method described in the above paragraph when said program is run on a computer.

According to yet another aspect of the invention these objects are achieved by a computer readable medium, having a program recorded thereon, where the program is to make a computer perform the method described in the penultimate paragraph above.

According to an additional aspect of the invention the object is achieved by a database for storing index information relating to digitised entities, which have been generated according to the proposed method.

According to yet an additional aspect of the invention the object is achieved by a server/client system for searching for digitised entities in a data collection as initially described, which is characterised in that an index database in the server/client system is organised, such that index information contained therein, for a particular digitised entity comprises at least one rank parameter. The rank parameter is indicative of a degree of relevance for at least one distinctive feature with respect to the digitised entity.

The invention provides an efficient tool for finding highly relevant non-text material on the Internet by means of a search query formulated in textual terms. An advantage offered by the invention is that the web pages, or corresponding resources, where the material is located need not contain any text information to generate a hit.

This is an especially desired feature, in comparison to the known solutions, since in many cases the non-text material may be accompanied by rather laconic, but counter intuitive text portions.

A particular signature for each unique digitised entity utilised in the solution according to the invention makes it possible eliminate any duplicate copies of digitised entities in a hit list obtained by a search. Naturally, this further enhances the search quality.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is now to be explained more closely by means of preferred embodiments, which are disclosed as examples, and with reference to the attached drawings.

Figure 1 illustrates the generation of a first rank component in a proposed rank parameter according to an embodiment of the invention,

Figure 2 illustrates an enhancement of the first rank component according to an embodiment of the invention,

Figure 3 illustrates the generation of a second rank component in the proposed rank parameter according to an embodiment of the invention,

Figure 4 demonstrates an exemplary structure of a search result according to an embodiment of the invention,

Figure 5 shows a block diagram over a server/client system according to an embodiment of the invention, and

Figure 6 illustrates, by means of a flow diagram, an embodiment of the method according to the invention.

DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION

The invention aims at enhancing the relevancy of any distinctive features, for instance keywords, being related to digitised entities and thereby improving the chances of finding relevant entities in future searches. In order to achieve this objective, at least one rank parameter is allocated to each distinctive feature that is related to a digitised entity. The embodiment of the invention described below refers to digitised entities in the form

of images. However, the digitised entities may equally well include other types of entities that are possible to identify uniquely, such as audio files or video sequences. Moreover, the digitised entities may either constitute sampled representations of analogue signals or be purely computer-generated entities.

Figure 1 shows four copies $c_a - c_d$ of one and the same image n that are stored at different locations in a data collection, for instance in an internetwork, like the Internet. The identity of the image n can be assessed by means of a so-called image signature, which may be determined from a total sum of all pixel values contained in the image. A corresponding identity may, of course, be assessed also for an audio file or a video file.

The copies $c_a - c_d$ of the image n are logically grouped together in a cluster C_n. Each copy $c_a - c_d$ is presumed to be associated with at least one distinctive feature in the form of a keyword. Typically, the keywords are data that are not necessarily being shown jointly with the image. On the contrary, the keywords may be collected from data fields which are normally hidden to a visitor of a certain web page. Thus, the keywords may be taken from HTML-tags such as *Meta*, *Img* or *Title* (HTML = HyperText Mark-up Language).

In this example a first copy c_a of the image n is associated with the keywords k_1, k_2, k_3, k_4 up to k_{ja}, a second copy c_b is associated with the keywords k_3, k_4, k_7, k_{19} up to k_{jb}, a third copy c_c is associated with the keywords k_1, k_3, k_4, k_5 up to k_{jc}, and a fourth copy c_d is associated with the keywords k_2, k_4, k_9, k_{12} up to k_{jd}. In order to determine the relevance of a particular keyword, say k_3, with respect to the image n a first rank component $\Gamma_n(k_3)$ is calculated according to the expression:

$$\Gamma_n(k_3) = \frac{\sum_i k_{i,3}}{|C_n|}$$

where $\sum_i k_{i,3}$ represents a sum of all occurrences of the keyword k_3 in the cluster C_n and $|C_n|$ denotes a total number of keywords in the cluster C_n, i.e. the sum of unique keywords plus any copies of the same.

However, it is also quite common that a particular keyword, for instance k_3, is associated with many *different* images. This is illustrated in figure 2. Here, a first cluster C_1 contains nine copies c_{11} - c_{19} of a first image n_1, a second cluster C_2 contains four copies c_{21} - c_{24} of a second image n_2 and a third cluster C_3 contains one copy c_{31} of a third image n_3. The keyword k_3 occurs twice (affiliated with c_{11} and c_{12}) in the first cluster C_1, three times (affiliated with c_{21}, c_{22} and c_{24}) in the second cluster C_2 and once (affiliated with c_{31}) in the third cluster C_3. The copy c_{12} occurs twice in the first cluster C_1, on one hand associated with the keyword k_3 and on the other hand associated with a different keyword. In both cases, however, it is the same image.

The first rank component Γ for the keyword k_3 may now be improved by means of a figure reflecting the strength in linkage between the keyword k_3 and the images n_1 - n_3 (or clusters C_1 - C_3) to which it has been associated. The keyword k_3 appears to have its strongest link to the second image n_2, since it is associated with the largest number of copies of this image, namely c_{21}, c_{22} and c_{24}. Correspondingly, the keyword k_3 has a second strongest link to the first image n_1 (where it occurs in two out of nine copies), and a third strongest link to the third image n_3. A normalisation with respect to the largest cluster (i.e. the cluster which includes the most copies) may be used to model this aspect. In this example, the largest cluster C_1 includes nine copies c_{11} - c_{19}. Therefore, a normalisation of the keyword k_3 with respect to the images n_1 - n_3 is obtained by multiplying the first rank component $\Gamma_n(k_3)$ with the respective number of occurrences in each cluster C_1 - C_3 divided by nine. Of course, the general expression becomes:

$$\Gamma_n(k_j) = \frac{\sum_i k_{i,j}}{|C_n|} \cdot \frac{|C_n|}{|C_{max}|} = \frac{\sum_i k_{i,j}}{|C_{max}|}$$

where $|C_{max}|$ is the largest number of keywords in a cluster for any image that includes the relevant keyword k_j, for instance k_3.

The first rank component Γ is made more usable for automated processing if it is also normalised, such that the highest first rank component Γ for a particular keyword is equal to 1. This is accomplished by dividing the expression above with the following denominator:

$$\frac{(\sum_i k_{i,j})max,k_j}{|C_{max}|}$$

where $(\sum_i k_{i,j})max,k_j$ denotes the number of occurrences of the keyword k_j in the cluster, which includes most occurrences of this keyword k_j. For instance, $(\sum_i k_{i,3})max,k_3$ is equal to 3 in the present example, since the keyword k_3 occurs most times in the second cluster C_2, namely three times.

Hence, the first rank component $\Gamma_n(k_j)$ for an image n with respect to keyword k_j is preferably modelled by the simplified expression:

$$\Gamma_n(k_j) = \frac{\sum_i k_{i,j}}{(\sum_i k_{i,j})max,k_j}$$

where $\sum_i k_{i,j}$ represents the sum of all occurrences of the keyword k_j in the cluster C_n and $(\sum_i k_{i,j})max,k_j$ is the number of occurrences of the keyword k_j in the cluster, which includes most occurrences of this keyword k_j.

However, in order to the improve the search performance in a database containing indexed elements referring to the digitised

entities, it is preferable to build an inverted index on keywords, such that a set of first rank components Γ is instead expressed for each keyword k_j. Thus, according to a preferred embodiment of the invention, the format of the first rank component is $k_j:\{\Gamma_n\}$. Consequently, the keyword k_3 in the example above obtains the following set of first rank components:

k_3: $\{\Gamma_2=1; \Gamma_1=2/3; \Gamma_3=1/3\}$

The first rank component $\Gamma_n(k_j)$ itself constitutes a fair reflection of the relevancy of a keyword k_j with respect to the image n. However, a more accurate figure can be obtained by combining the first rank component $\Gamma_n(k_j)$ with a proposed second rank component $\Pi_n(k_j)$, which will be described below.

Figure 3 illustrates how the second rank component $\Pi_n(k_j)$ may be generated according to an embodiment of the invention.

A digitised entity, e.g. an image 301, is presumed to be associated with distinctive features k_1, k_2 and k_3, for instance in the form of keywords, which are found at various positions P in a descriptive field F. Each distinctive feature k_1 - k_3 is estimated to have a relevance with respect to the digitised entity 301 that depends on the position P in the descriptive field F in which it is found. A weight factor w_1 – w_p for each position 1 – p in the descriptive field F reflects this. In the illustrated example, a first distinctive feature k_1 and a second distinctive feature k_2 are located in a position 1 in the descriptive field F. Both the distinctive feature k_1 and the distinctive feature k_2 occur a number η_1 times in this position. There are no distinctive features in a second position 2. However, various distinctive features may be located in following positions 3 to p-2 (not shown). The field F contains η_2 copies of the first distinctive feature k_1 in a position p-1 and η_1 copies of the second distinctive feature k_2 respective η_3 copies of a third distinctive feature k_3 in a position p.

Hence, depending on the position 1 – p in which a certain distinctive feature k_1 - k_3 is found, the distinctive feature k_1 - k_3 is allocated a particular weight factor w_1 – w_p. Furthermore, a relevancy parameter s_1 – s_4 is determined for every distinctive feature k_1 - k_3, which depends on how many times η_1 - η_3 the distinctive feature k_1 - k_3 occurs in a position 1 - p relative a total number of distinctive features in this position 1 – p.

Thus, both the first distinctive feature k_1 and the second distinctive feature k_1 obtain the same relevancy parameter s_1, which can be calculated as $s_1 = \eta_1/(2\eta_1) = 1/2$ in the position 1. This parameter s_1 is further weighted with a weight factor w_1 in respect of the digitised entity 301. The same calculations are performed for all the positions 2 – p to obtain corresponding relevancy parameters s_1 – s_4 for these positions.

Alternatively, the relevancy parameter s_P can be determined as $s_P(k_{j\neq i})=1-\gamma\sum_i k_i$, where $\gamma\sum_i k_i$ represents a "penalty" that decreases the relevancy for a distinctive feature k_j in a position P, for each distinctive feature in this position, which is different from the distinctive feature k_j. Naturally, other alternative formulas for determining the relevancy parameter s_P are also conceivable.

Nevertheless, a combined measure is determined, which fully captures the relationship between distinctive features k_j and digitised entities n. The expression:

$$\Pi(n, k_j) = \frac{\sqrt{\sum_{i=1}^{p}(w_i \cdot s_{i,j})^2}}{\sqrt{\sum_{i=1}^{p} w_i^2}}$$

constitutes a reflection of the relevance of a distinctive feature k_j with respect to a particular digitised entity n. The variable w_i denotes the weight factor for a position i and the variable $s_{i,j}$

denotes the relevancy parameter for a distinctive feature k_j in the position i.

In analogy with the first rank component Γ, its is preferable also to normalise and build an inverted index on keywords. The second rank component Π is thus given a format $k_j:\{\Pi_n\}$, where the first component Π_n for a particular distinctive feature k_i is always equal to 1.

Table 1 below shows an explicit example over weight factors w_i for a certain positions P in a descriptive field F related to an image.

Position (P)	Field (F)	Weight factor (w_P)
1	pageSite	50
2	pageDir	40
3	pageName	50
4	pageTitle	80
5	pageDescription	90
6	pageKeywords	90
7	pageText	20
8	imageSite	50
9	imageDir	60
10	imageName	100
11	imageAlt	100
12	imageAnchor	80
13	imageCenterCaption	90
14	imageCellCaption	90
15	imageParagraphCaption	90

Table 1

According to an embodiment of the invention, the second rank component $\Pi_n(k_j)$ is used as an alternative to the first rank component $\Gamma_n(k_j)$. The second rank component $\Pi_n(k_j)$ is namely

also a per se good descriptor of the relevancy of a keyword k_j with respect to the image n.

In a preferred embodiment of the invention, however, the first rank component Γ and the second rank component Π are merged into a combined rank parameter Δ according to the expression:

$$\Delta = \sqrt{\frac{(\alpha\Gamma)^2 + (\beta\Pi)^2}{\alpha^2 + \beta^2}}$$

where α is a first merge factor and β is a second merge factor. For instance, $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$. However, any other range of the merge factors α; β are likewise conceivable.

Finally and in similarity with the first and second rank components Γ and Π respectively, it is preferable to normalise and build an inverted index on keywords, such that it obtains a format $k_j:\{\Delta_n\}$, where the first component Δ_n for a particular distinctive feature k_j is always equal to 1.

When all, or at least a sufficiently large portion, of the digitised entities in the data collection have been related to at least one distinctive feature and a corresponding rank component/ parameter (Γ, Π or Δ), an index database is created, which also at least includes a field for identifying the respective digitised entity and a field containing one or more locators that indicate where the digitised entity can be retrieved. Moreover, it is preferable if the index database contains an intuitive representation of the digitised entity. If the digitised entity is an image, a thumbnail picture constitutes a suitable representation. If, however, the digitised entity is an audio file or multimedia file, other representations might prove more useful, for instance in the form of logotypes or similar symbols.

Figure 4 demonstrates an exemplary structure of a search result according to an embodiment of the invention. The search result is listed in a table 400, where a first column E contains the

identity $ID_1 - ID_m$ of the entities that matched the search criteria sufficiently well. A second column K contains an inventory of ranked distinctive features $\Delta(k_1)$ - $\Delta(k_{23})$ for each digitised entity. A third column R includes a characterising representation (or an illustrative element) $r_1 - r_m$ of the entity and a fourth column L contains at least one locator $l_1 - l_m$ to a corresponding "full version" of the entity. In case the data collection is an internetwork the locator $l_1 - l_m$ is typically a URL (Universal Resource Locator). However, any other type of address is equally well conceivable.

Naturally, the search result structure may also include arbitrary additional fields. A reduced set of fields may then be presented to a user. It could, for instance, be sufficient to display only the representation $r_1 - r_m$ and / or a limited number of the distinctive features, with or / without their respective ranking.

Figure 5 shows a block diagram over a server/client system according to an embodiment of the invention, through which data may be both indexed, searched and retrieved. Digitised entities are stored in large and rather unstructured a data collection 510, for instance the Internet. An indexing input device 520 gathers information ID_n, {K}; L from the data collection 510 with respect to digitised entities contained therein. The information ID_n, {K}; L includes at least an identity field ID_n that uniquely defines the digitised entity E, a set of distinctive features {K} and a locator L. Additional data, such as file size and file type may also be gathered by the indexing input device 520. It is irrelevant exactly how the information ID_n, {K}; L is entered into the indexing input device 520. However, according to a preferred embodiment of the invention, an automatic data collector 521, for instance in the form of an web crawler, in the indexing input device 520 regularly accumulates C the information ID_n, {K}; L as soon as possible after addition of new items or after updating of already stored items. An index generator 522 in the indexing input device 520 creates index information I_E on basis of the information ID_n, {K}; L according to

the methods disclosed above. An index database 530 stores the index information I_E in a searchable format, which is at least adapted to the operation of a search engine 540.

One or more user client terminals 560 are offered a search interface towards the index information I_E in the index database 530 via a user client interface 550. A user may thus enter a query phrase Q, for instance, orally via a voice recognition interface or by typing, via a user client terminal 560. Preferably, however not necessarily, the user client interface 550 re-formulates the query Q into a search directive, e.g. in the form of a search string S, which is adopted to the working principle of the search engine 540. The search engine 540 receives the search directive S and performs a corresponding search S' in the index database 530.

Any records in the database 530 that match the search directives S sufficiently well are sorted out and returned as a hit list {H} of digitised entities E to the user client interface 550. If necessary, the user client interface 550 re-formats the hit list {H} into a search result having a structure H(R, L), which is better suited for human perception and / or adapted to the user client terminal 560. The hit list {H} preferably has the general structure shown in figure 4. However, the search result H(R, L) presented via the user client terminal 560 may have any other structure that is found appropriate for the specific application. If the query phrase Q comprises more than one search term (or distinctive feature), the search result H(R, L) has proven to demonstrate a desirable format when each search term in the hit list {H} is normalised before presentation to the user, such that a first combined rank parameter Δ_n for each search term is equal to 1. For instance, a hit list {H} resulting from a search query Q = "ferarri 550" is normalised such that the first combined rank parameter Δ_n = 1 both with respect to "ferarri" and with respect to "550". Any additional combined rank parameters Δ_m for the respective search terms may, of course, have arbitrary lower value depending on the result of the search.

The signature associated with each unique digitised entity makes it possible eliminate any duplicate copies of digitised entities in the search result H(R, L). Such elimination produces a search result H(R, L) of very high quality and relevance.

A minimum requirement is that the data sent to the user client terminal 560 includes a characteristic representation R of the digitised entities in the hit list {H} and corresponding locators L, e.g. URL, for indicating at least one storage location in the data collection 510. The latter gives the user at least a theoretical possibility to retrieve full versions of the digitised entities. In practice, however, the retrieval may be restricted in various ways, for instance by means of copyright protection and therefore require the purchase of the relevant rights.

The units 510 - 560 may either be physically separated from each other or be co-located in arbitrary combination.

In order to sum up, a method of generating a searchable index for digitised entities according to an embodiment of the invention will now be described with reference to a flow diagram in the figure 6.

A first step 601 involves input of basic information that contains one or more distinctive features being related to digitised entities in a data collection. A following step 602 creates rank parameters for each of the digitised entities on basis of the input information. Then, a step 603 generates a searchable index for the rank parameters and finally, the searchable index is stored in a searchable database in a step 604.

All of the process steps, as well as any sub-sequence of steps, described with reference to the figure 6 above may be controlled by means of a computer program being directly loadable into the internal memory of a computer, which includes appropriate software for controlling the necessary steps when the program is run on a computer. The computer program can likewise be recorded onto arbitrary kind of computer readable medium.

The term "comprises/comprising" when used in this specification is taken to specify the presence of stated features, integers, steps or components. However, the term does not preclude the presence or addition of one or more additional features, integers, steps or components or groups thereof.

The invention is not restricted to the described embodiments in the figures, but may be varied freely within the scope of the claims.

Claims

1. A method of indexing digitised entities (E) in a data collection (510) comprising:

inputting basic information (ID_n, {K}, L) pertaining to at least one distinctive feature ({K}) and at least one locator (L) for each digitised entity (E) in a set of entities from the data collection (510),

generating searchable index information (I_E) related to the digitised entities (E) in the set on basis of the basic information (ID_n, {K}, L), and

storing the index information (I_E) in an index database (530), **characterised by**

generating the index information (I_E) for a particular digitised entity (E: ID_n) on basis of at least one rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$) derived from the basic information (ID_n, {K}, L), the at least one rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$) being indicative of a degree of relevance for at least one distinctive feature (k_3, k_5; k_{19}) with respect to the digitised entity (E: ID_n).

2. A method according to claim 1, **characterised by** the at least one rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$) being based on a first rank component (Γ) that is generated by a first algorithm, which involves ranking individual distinctive features ($k_1 - k_{jd}$) related to the digitised entity (E: n) on basis of a relative occurrence of the individual distinctive features ($k_1 - k_{jd}$) with respect to one or more copies ($c_a - c_d$) of the digitised entity (E: n) in the data collection (510).

3. A method according to claim 1, **characterised by** the first algorithm involving the following steps, with respect to a particular distinctive feature (k_3), for the digitised entity (E: n):

grouping at least one copy (c_a–c_d, c_{11}–c_{19}, c_{21}–c_{24}; c_{31}) of at least the digitised entity (E: n; n_1, n_2, n_3) in a cluster (C_n), each cluster (C_n, C_1, C_2, C_3) containing one or more copies of the same digitised entity (E: n; n_1, n_2, n_3),

counting a total number of occurrences of the particular distinctive feature (k_3) in each cluster (C_n, C_1, C_2, C_3), and

calculating a ratio between the total number of occurrences of the particular distinctive feature (k_3) in the cluster (C_n) for the digitised entity (E: n) and the total number of occurrences of the particular distinctive feature (k_3) in a cluster (C_2) which includes a largest number of the particular distinctive feature (k_3).

4. A method according to any one of the claims 1 - 3, **characterised by** the at least one rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$) being based on a second rank component (Π) that is generated by a second algorithm, which involves ranking at least one individual distinctive feature (k_1, k_2, k_3) related to the digitised entity (301) on basis of a position (P) of the least one individual distinctive feature (k_1, k_2, k_3) in a descriptive field (F) associated with the digitised entity (E).

5. A method according to claim 4, **characterised by** generating the second rank component (Π) on basis of a particular weight factor ($w_1 - w_p$) being linked to each position (1 – p) in the descriptive field (F), the weight factors ($w_1 - w_p$) reflecting a distinctive feature's (k_1, k_2, k_3) significance with respect to its position (P) in the descriptive field (F).

6. A method according to claim 5, **characterised by** generating the second rank component (Π) on basis of a relevancy parameter ($s_1 - s_4$) reflecting a distinctive feature's (k_1) significance in relation to other distinctive features (k_2) in a particular position (p) in the descriptive field (F).

7. A method according to any one of the claims 4 - 6, **characterised by** the generating of the rank parameter ($\Delta(k_3)$,

$\Delta(k_5)$; $\Delta(k_{19})$) involving a combination of the first rank component (Γ) with the second rank component (Π).

8. A method according to claim 7, **characterised by** combining the first rank component (Γ) with the second rank component (Π) according to the expression:

$$\sqrt{\frac{(\alpha\Pi)^2 + (\beta\Gamma)^2}{\alpha^2 + \beta^2}}$$

where Γ represents the first rank component, Π represents the second rank component, α represents a first merge factor and β represents a second merge factor.

9. A method according to any one of the preceding claims, **characterised by** the digitised entities (E) including at least one of the file types: a text document, an image, a video sequence and an audio sequence.

10. A method according to claim 9, **characterised by** at least one of the digitised entities (E) constituting a sampled representation of an analogue signal.

11. A method according to claim 9, **characterised by** at least one of the digitised entities (E) constituting a computer generated entity.

12. A method according to any one of the preceding claims, **characterised by** the distinctive feature (k_1 - k_{jd}) being a keyword.

13. A computer program directly loadable into the internal memory of a digital computer, comprising software for

performing the steps of any of the claims 1–12 when said program is run on a computer.

14. A computer readable medium, having a program recorded thereon, where the program is to make a computer perform the steps of any of the claims 1–12.

15. A database for storing index information (I_E) relating to digitised entities (E), which have been generated according to any one of the claims 1–12.

16. A server/client system for searching for digitised entities (E) in a data collection (510) comprising

an indexing input device (520) for collecting basic information (ID_n, {K}, L) pertaining to at least one distinctive feature ({K}) and at least one locator (L) for each digitised entity (E) in a set of entities from the data collection (510),

an index database (530) for storing index information (I_E) relating to the digitised entities (E) in the set,

a search engine (540) for receiving search directives (S) and in response thereto performing searches (S') in the index database (530), and

a user client interface (550) for receiving a search request (Q) from at least one user client terminal (560), forwarding the search request (Q) as a search directive (S) to the search engine (540), receiving a hit list ({H}) of digitised entities (E) and returning a result ((H(R, L)) of a corresponding search (S') in the index database (530) to the at least one user client terminal (560), **characterised in that**

the index database (530) is organised such that the index information (I_E) for a particular digitised entity (E: ID_n) comprises at least one rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$), which is indicative of a degree of relevance for at least one distinctive feature (k_3, k_5; k_{19}) with respect to that digitised entity (E: ID_n).

17. A server/client system according to claim 16, **characterised in that** the indexing input device (520) includes an index generator (522) for receiving the basic information (ID_n, {K}, L) and producing in response thereto the at least one rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$).

18. A server/client system according to any one of the claims 16 or 17, **characterised in that** the rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$) includes a first rank component (Γ) indicating a ranking of at least one individual distinctive feature ($k_1 - k_{jd}$) related to the digitised entity (E: n) on basis of a relative occurrence of the at least one individual distinctive feature ($k_1 - k_{jd}$) with respect to one or more copies ($c_a - c_d$) of the digitised entity (E: n) in the data collection (510).

19. A server/client system according to any one of the claims 16-18, **characterised in that** the rank parameter ($\Delta(k_3)$, $\Delta(k_5)$; $\Delta(k_{19})$) includes a second rank component (Π) indicating a ranking of at least one individual distinctive feature (k_1, k_2) related to the digitised entity (301) on basis of

a position (P) of the least one individual distinctive feature (k_1, k_2) in a descriptive field (F) associated with the digitised entity (E), and

a relevancy parameter ($s_1 - s_4$) reflecting a distinctive feature's (k_1) significance in relation to other distinctive features (k_2) in a particular position (p) in the descriptive field (F).

20. A server/client system according to any one of the claims 16-19, **characterised in that** the indexing input device (520) includes an automatic data collector (521) for finding relevant digitised entities (E) in the data collection (510) and creating there from the set of entities.

21. A server/client system according to any one of the claims 16-20, **characterised in that** each of the digitised entities (E) in the hit list ({H}) is associated with an identifier (ID_1 – ID_m), at least one rank parameter ($\Delta(k_2)$, $\Delta(k_5)$; $\Delta(k_6)$ - $\Delta(k_5)$; $\Delta(k_{12})$) and at least one locator (l_1 – l_m) for indicating a storage location in the data collection (510).

22. A server/client system according to claim 21, **characterised in that** each of the digitised entities (E) in the hit list ({H}) is also associated with an illustrative element (r_1 – r_m) to be displayed on the user client terminal (560) together with the respective digitised entity (E).

23. A server/client system according to claim 22, **characterised in that** the illustrative element (r_1 – r_m) is a thumbnail picture.

24. A server/client system according to any one of the claims 20–23, **characterised in that** the data collection (510) is an internetwork and the indexing input device (520) includes a web crawler (522).

25. A server/client system according to any one of the claims 16–24, **characterised in that** the digitised entities (E) include at least one of the file types: a text document, an image, a video sequence and an audio sequence.

26. A server/client system according to claim 25, **characterised in that** the digitised entities (E) are stored in at least one of the formats: AIF, AIFC, AIFF, AU, AVI, BMP, DIVX, DOC, EPS, GIF, ICO, JPEG, JPG, MOV, MP3, MP4, MPEG, MPEG4, MPG, PDF, PNG, PPT, PS, QT, RA, RAM, RAS, SND, TIF, TIFF, VCD, WAV, XLS and XMP.

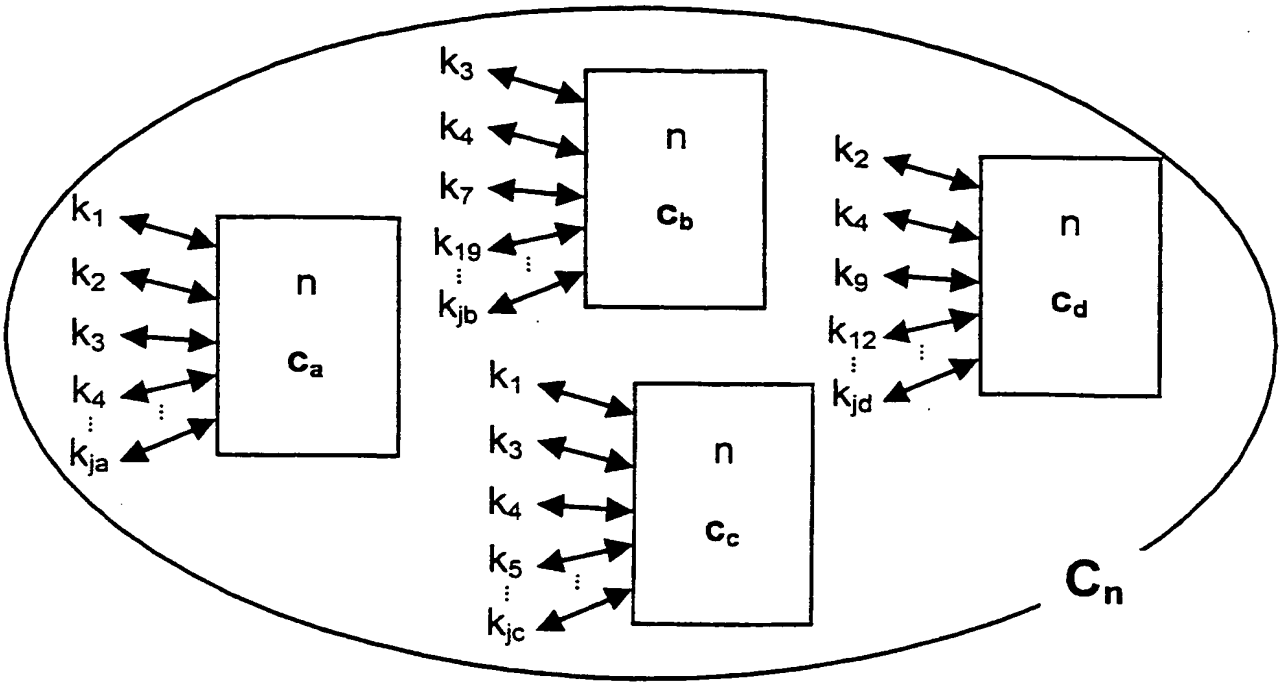

k1
k2
k3
k4
kja
n
ca
k3
k4
k7
k19
kjb
n
cb
k1
k3
k4
k5
kjc
n
cc
k2
k4
k9
k12
kjd
n
cd
Cn

Fig. 1

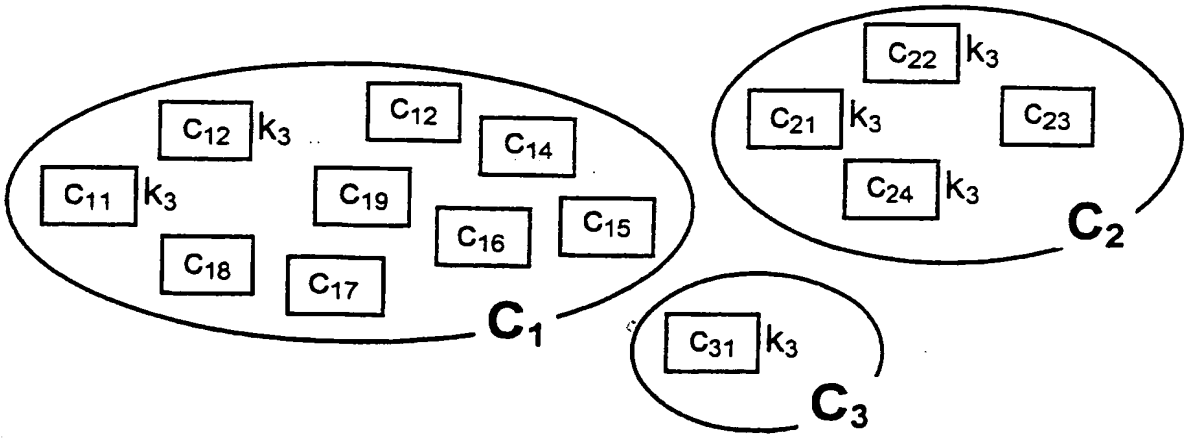

C12 k3
C12
C14
C11 k3
C19
C16
C15
C18
C17
C1
C22 k3
C21 k3
C23
C24 k3
C2
C31 k3
C3

Fig. 2

F	N	S	P
k_1	η_1	s_1	1
k_2	η_1	s_1	1
---	---	---	2
⋮	⋮	⋮	⋮
k_1	η_2	s_1	p-1
k_2	η_1	s_4	p
k_3	η_3	s_3	p

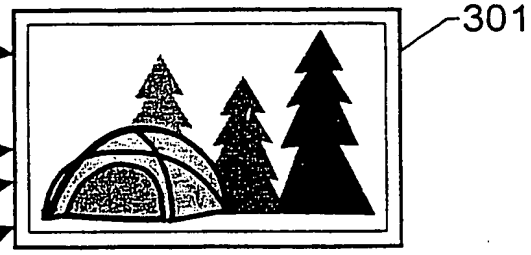


Fig. 3

E	K	R	L
ID_1	$\Delta(k_2), \Delta(k_5), \Delta(k_6)$	r_1	l_1
ID_2	$\Delta(k_1), \Delta(k_{23})$	r_2	l_2
⋮	⋮	⋮	⋮
ID_n	$\Delta(k_3), \Delta(k_5), \Delta(k_{19})$	r_n	l_n
⋮	⋮	⋮	⋮
ID_m	$\Delta(k_5), \Delta(k_{12})$	r_m	l_m

400

Fig. 4

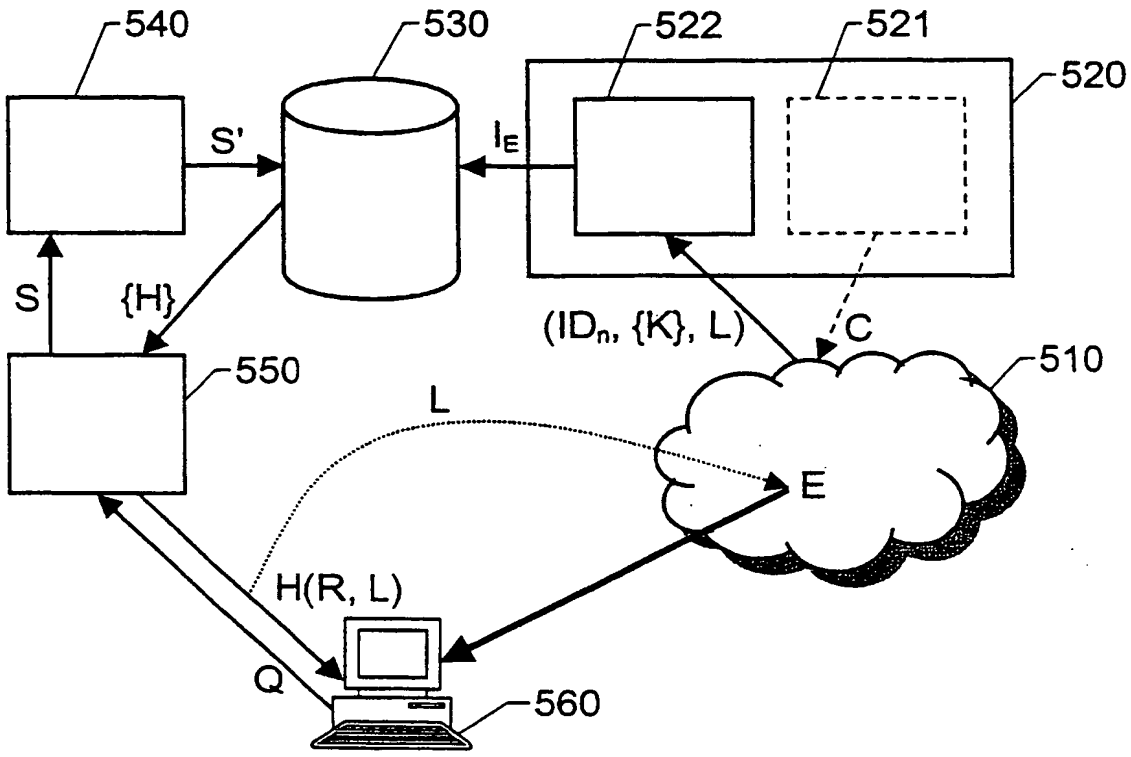

540
530
522
521
520
S'
I_E
S
{H}
$(ID_n, \{K\}, L)$
C
550
510
L
E
H(R, L)
Q
560

Fig. 5

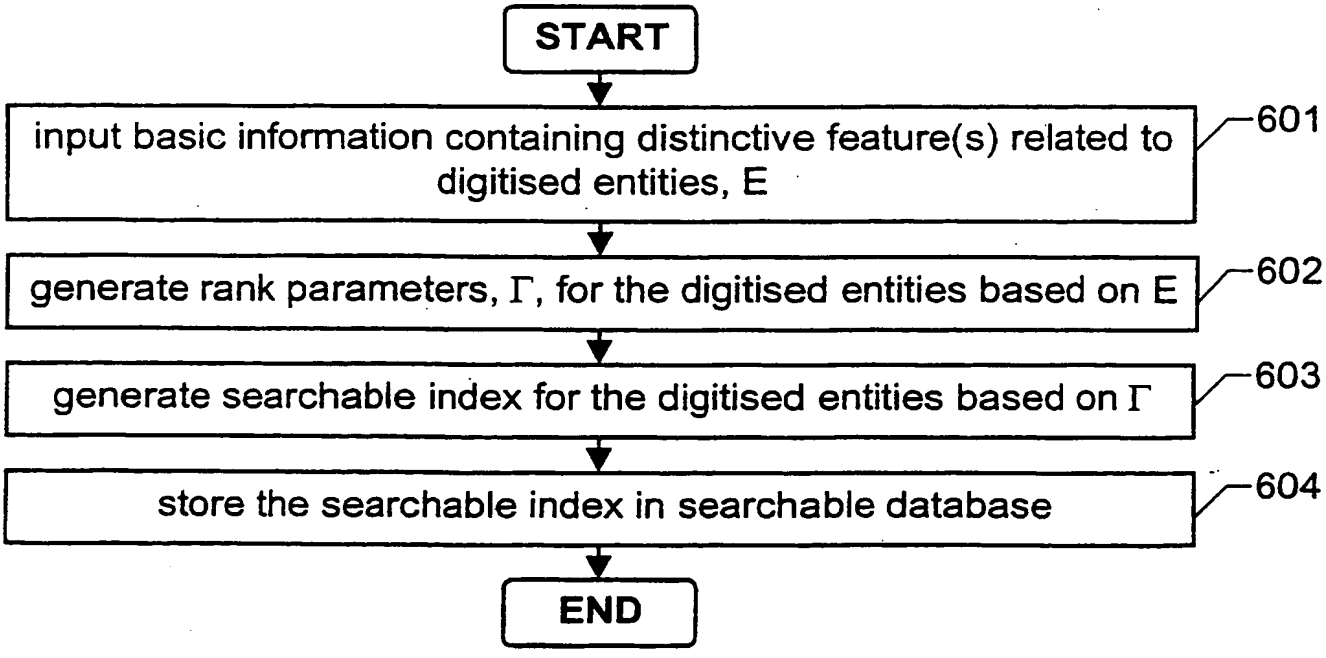

START
input basic information containing distinctive feature(s) related to digitised entities, E
601
generate rank parameters, Γ, for the digitised entities based on E
602
generate searchable index for the digitised entities based on Γ
603
store the searchable index in searchable database
604
END

Fig. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 02/00462

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: G06F 17/30
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6084595 A (BACH, J.R. ET AL), 4 July 2000 (04.07.00), column 2, line 12 - column 3, line 25, figure 3, abstract --	1-26
A	EP 0596247 A2 (MOTOROLA, INC.), 11 May 1994 (11.05.94) --	1-26
A	WO 0033575 A1 (YUEN, H.), 8 June 2000 (08.06.00) -- --------	1-26

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 June 2002	25 -06- 2002
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Oskar Pihlgren /OGU Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

Information on patent family members

01/05/02

International application No.
PCT/SE 02/00462

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6084595 A	04/07/00	NONE	
EP 0596247 A2	11/05/94	JP 6282588 A	07/10/94
WO 0033575 A1	08/06/00	AU 2034600 A	19/06/00
		EP 1142329 A	10/10/01

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 02/00462

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

The application according to the independent claims 1, 16 is considered not to involve an inventive step. The application therefore lacks unity a posteriori since many of the dependent claims that are referring directly to an independent claim do not have any special technical features in common. However the entire application has still been searched without effort justifying any additional fees.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☒ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

This Page Blank (uspto)